recerTIC UPC: a new approach to bibliometric analysis for a research university

Rubèn Pocull[1], Miquel Codina-Vila[1], Ruth Inigo[1], Sara Matheu[1], Andrés Pérez[1], Javier Clavero[2]

[1]*ruben. pocull@upc. edu; miquel. codina@upc. edu; ruth. inigo@upc. edu; sara. matheu@upc. edu; andres. perez@upc. edu*
Polytechnic University of Catalonia. Rector Gabriel Ferraté Library, C / Jordi Girona 1-3, 08034, Barcelona (Spain)

[2]*javier. clavero@upc. edu*
Polytechnic University of Catalonia. Libraries, Publications and Archives Service, C / Jordi Girona 1-3, 08034, Barcelona (Spain)

Introduction

The Biblioteca Rector Gabriel Ferraté (Rector Gabriel Ferraté Library, BRGF) of the Universitat Politècnica de Catalunya (Polytechnic University of Catalonia, UPC) presents *recerTIC UPC* <https://tinyurl.com/y4os625c>: a set of ten bibliometric works on the same number of subjects, which is designed to generate a meaningful and easily understandable map of the UPC's scientific publications on topics relevant to the field of Information and Communications Technologies (ICT). These studies place emphasis on showing the links between internal and external collaborations of researchers who author these publications, as well as the dynamics of transversality between the subject areas of the analyzed research publications.

Methodology

The studies are based on journal articles and conference publications published by UPC professors between 2007 and 2017.

The process of creating *recerTIC UPC* began with the directors of ICT learning centers petitioning for bibliometric studies on specific subject areas that had previously been left out of bibliometric studies prepared by the BRGF.

These subjects were determined in close collaboration with UPC researchers specialized in the corresponding areas. Consideration was given to aspects such as the relevance and emergence of each technology, its social impact, its future potential, and its strategic relevance for the UPC. Ultimately, the areas chosen were: 5G, Computer Security, Embedded Systems, Machine Learning, Smart Sensors, Bioinformatics, Data Science and Engineering, IoT, Robotics and Vehicle-to-Everything.

The database chosen as the primary source of information was Web of Science Core Collection (WoS), by Clarivate Analytics. In choosing this, the definition, scope and level of granularity for the WoS *Subject Categories* were taken into account. This is appropriate for the subsequent analysis not only of the relationships between subject nodes, but also of the inclusion of keywords and subject descriptors in most registries.

The references of the WoS publications were extracted by using non-controlled vocabulary in the WoS field *TS=Topic*. This field tag retrieves terms within the fields *Title, Abstract, Author Keywords* and *Keywords Plus®*.

Choosing keywords was the most critical part for obtaining reliable results. Their selection relied on the collaboration of relevant researchers who identified concepts and terms that define the contents of each area of study. These terms may be of a generic scope and designate various disciplines, techniques and methodologies that are closely related to the research field being considered.

Paradigmatically for the fields that form the theoretical basis of many *recerTIC UPC* subjects (for example, Statistics, Mathematics and Physics), the difficulty lies in recovering publications relevant to the study subject when this subject is often tacit. In order to minimize this problem, we have attempted to ensure that the terms related to the most basic and methodological aspects of each study do not distort or add noise to the results.

However, it was necessary on various occasions to iterate the algorithms so that they would introduce concepts that were not included in the first approaches, which was done in order to obtain balanced and representative results.

Generating maps with GenMap

The interactive node maps were created with GenMap, an application that was custom-designed for the ITC staff at the Libraries, Publications and Archives Service. We used the library cytoscape.js, which specializes in representing graphs of data, nodes and edges – all of which are extracted from various information sources (currently: WoS, Scopus, and DBLP, among others) – with the aim of

facilitating visualization while making it more interactive.

Once the map is created, the next step is to import the extracted file from the database corresponding to the information we want to view. Parameterizing the concepts for representation allows generating a graph in which the nodes indicate the analyzed concept, its identifying title, the size according to a chosen criterion and a color representing the characteristic intended to be emphasized. Edges connect the nodes and represent the relationships between them.

Once the data entry process has been completed, the visualization module allows all users to view the knowledge map and interact with it.

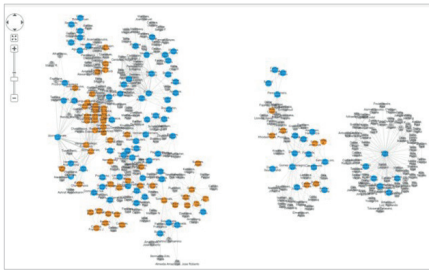

Figure 1. Interactive node map created with GenMap

The user can adjust the visualization by means of the following: zoom; moving the nodes; searching nodes on the map with a drop-down menu; selecting and highlighting nodes and/or edges; viewing node and edge information; modifying the map's graphical settings; exporting the map in various formats (with the possibility of inserting it into a web page); and, finally, sharing the map via social networks.

Results

The results obtained by the *recerTIC UPC* studies show data and significant information regarding: the location and analysis of clusters and communities of collaboration and co-authorship; relationships between subject areas; and other aspects pertaining to the evolution of UPC scientific production in the areas analyzed.

The co-author maps show different clusters of authors in each subject area. Diverse clusters exist because various research groups take different approaches to working in this area in separate stages of development. Information is also collected on production that forms a part of the methodology or the practical application of the research in another field (for example, machine learning applied to fish farm management or 5G applied to traffic management).

Distinguishing the nodes by color reveals several dynamics regarding internal co-authorship, co-authorship with related research centers, and co-authorship that is external to the UPC.

Regarding the node maps corresponding to the WoS categories assigned to the publications, these indicate the weight of each category and how they relate to each other within the framework of the subject area studied.

Conclusions

The aim of *recerTIC UPC* is to analyze and highlight the power of our university's scientific production in certain areas covering the most relevant research on current technology. To achieve this, we generated a "satellite image" of research at the UPC, which can be browsed as a set and then "downloaded" in order to discover details and new niches for investigation, as well as for applying and/or transferring the research.

In addition, *recerTIC UPC* constitutes a step forward in our work with bibliometric data. Some of the characteristics that define this new generation of our bibliometric productions are: the use of new technologies; a presentation that facilitates analyzing results; and options for interacting with the data that we have processed through our searches.

References

Codina-Vila, M., & Íñigo, R. (2015). De la investigación al investigador. Adaptando servicios en la Biblioteca Rector Gabriel Ferraté. *El profesional de la información, 24* (5), 648-655. doi:10.3145/epi.2015.sep.13

Falagas, M. E., Pitsouni, E. I, Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal 22*(2), 338-342. doi:10.1096/fj.07-9492LSF

Glänzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In Henk F. Moed, Wolfgang Glänzel, & Ulrich Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257-276). Dordrecht: Springer.

Jansen, B. J., & Rieh, S. Y. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology 61*(8), 1517-1534. doi:10.1002/asi.21358

Thelwall, M. (2008). Bibliometrics to webometrics. *Journal of Information Science 34* (4), 605-621. doi:10.1177/0165551507087238